Цао Паньпань
Аспирант
Уральский федеральный университет имени первого
Президента России Б.Н. Ельцина
Россия, г. Екатеринбург
Научный руководитель: Корнеева Лариса Ивановна
проф., д-р пед.н.

АКТУАЛЬНОЕ ОСТОЯНИЕ ИССЛЕДОВАНИЙ В ОБЛАСТИ ПЕРЕВОДОВЕДЕНИЯ В РОССИИ В УСЛОВИЯХ ЦИФРОВОЙ БАЗЫ ЯЗЫКОВЫХ ДАННЫХ

Аннотация. *Цель данной статьи - доказать, что расширение объема некоторых видов переводческой деятельности не обязательно ведет к новым подходам и концепциям в переводческой науке. В настоящее время наиболее оптимальной методологией изучения перевода является сочетание традиционной парадигмы эквивалентности и коммуникативно-функционального подхода к переводу.*

Ключевые слова: *корпусная лингвистика, цифровая база языковых данных, машинное обучение переводу.*

Tsao Panpan
Postgraduate student
Ural Federal University
Russia, Ekaterinburg
Academic supervisor: Korneeva Larisa Ivanovna

CURRENT STATUS OF RESEARCH IN THE FIELD OF TRANSLATION IN RUSSIA IN THE CONDITIONS OF A DIGITAL LANGUAGE DATABASE

Abstract. *The aim of this paper is to prove that the expansion of the scope of certain translation activities does not necessarily lead to new approaches and concepts in translation studies. Currently, the most optimal methodology for the study of translation is a combination of the traditional equivalence paradigm and the communicative-functional approach to translation.*

Keywords: *corpus linguistics, digital language database, machine learning translation*

Цифровые языковые данные представляют особый интерес в социальных науках, потому что из них мы можем получить обобщенные сведения о том, как люди живут, что ими движет, на какие группы они делятся, какие существуют социальные практики. Достаточно много примеров исследований, в которых для понимания этих социальных практик используются разные цифровые «следы», которые люди оставляют за собой. Это и их активность в соцсетях, и их предпочтения, покупки, передвижения по городу и так далее. Поскольку главной социальной практикой является язык, коммуникация с использованием естественного языка, то становится интересно, какие здесь могут быть возможности и, собственно говоря, где вообще взять эти большие данные, эти цифровые «следы», цифровые образцы, чтобы мы могли что-то понять про то, как язык используется, как он развивается, как люди говорят и как они не говорят.

К большому объему данных в переводоведении обращаться стали достаточно давно. Специалисты начали собирать корпуса, то есть такие большие наборы текстов, объединенные некоторой идеей, общей тематикой, которые в дальнейшем специальным образом обрабатывались, снабжались морфологической разметкой. И, таким образом, подготавливались ресурсы, для того чтобы потом ученые, специалисты могли к ним обращаться и получать определенные сведения, делать выборку данных и работать с ними.

С самого начала корпуса использовались в двух направлениях. С одной стороны, ученые-теоретики с помощью корпусов могли получить примеры

употребления определенной конструкции, примеры использования определенного класса глаголов. И конечно же, кроме самих примеров первое, что дает корпус, — это частотность их употребления. Частотность — это вообще самая главная вещь, которая эксплуатируется в корпусной лингвистике и в работе с большими или, может быть, не самыми большими, но все равно значительными языковыми данными.

Собственно говоря, с этой самой частотностью связано и второе направление использования корпусов — для решения задач компьютерного перевода, для машинного обучения. Когда стоит задача построения какой-то языковой модели, используется корпус, снабженный определенной разметкой. Разметка выделяет определенные интересующие классы элементов, и далее происходит разработка программы, специально написанного скрипта, который учится эти элементы различать. Таким образом, решается задача, связанная непосредственно с переводоведением. Например, задача морфологического анализа. Программу разрабатывают с целью отличить существительное от глагола. Каким образом? На вход алгоритма программы дается корпус, где слова размечены по видам: существительное, глагол, прилагательное, предлог. И дальше по разным свойствам появления того или иного тега высчитывается некоторая вероятность того, будет ли слово существительным, или глаголом, или какой-то другой частью речи. И алгоритм принимает это решение. Такие программы уже есть и для русского языка, и работают вполне хорошо.

Но кроме того, решаются самые разные задачи, связанные, например, с тематическим моделированием, то есть с определением темы текста. Или, более конкретно, с делением текстов по определенным тематикам. На вход подаются тексты с уже размеченной тематикой, на выходе получаются эти классы. Или на вход подаются тексты с какой-то другой разметкой, там тематика может быть не определена, но какие-то свойства все равно есть. И на выходе алгоритм машинного обучения делит этот корпус на какие-то тексты, близкие по тематике.

Такой традиционный анализ примеров, полученных из корпуса, и машинное обучение — это не тот анализ данных, который имеется в виду, когда

идет речь о больших данных, которые можно выявить только статистическим образом. Это направление стало развиваться недавно, и здесь есть интересное противоречие. Когда анализируются большие языковые данные, то задача либо очень сужается, и таким образом происходит сужение этих данных, и анализ остается в пределах переводоведения, либо существует риск выйти за рамки переводоведческого исследования. То есть, в данном случае идет речь уже не о языке, а о изменении жизни, которая отображена в текстах, подвергшихся анализу.

Другая большая история связана с корпусом, с ресурсом, который называется Google Ngram. Этот ресурс делали не лингвисты и переводчики, а биоинформатики. В 2011 году в журнале Science была опубликована статья, которая была названа «Quantitative analysis of culture using millions of digitized books». Как видно из названия, это совершенно не про лингвистику, а про культуру. Но факт в том, что эта статья была о том, как можно изучать культуру и язык с помощью анализа данных Google Books. Google Books — это как раз очень-очень-очень большие данные. И это как раз такие «следы», которые оставляют культура и естественный язык как главное средство передачи культуры в нашей текущей цифровой реальности. Потому что Google Books — это примерно 6% всех когда-либо опубликованных человечеством книг на восьми языках.

Главное достижение данных исследователей заключается в том, что они придумали, как можно предоставить миру, исследователям эти данные, не нарушая ничьи авторские права. Это стало огромным прорывом. Специалистами была создана ресурсная база, в которой исчислили все слова, а также сочетания слов от одного слова и до пяти слов подряд, которые встречаются в корпусе Google Books.

Кроме этого, конечно, были сделаны и общекультурные исследования. Самое знаменитое — исследование того, как была организована цензура в гитлеровской Германии и как упоминание еврейских деятелей культуры,

например Шагала, не меняется в англоязычных книгах, но резко падает в немецкоязычных.

Возвращаясь к переводоведению: оказывается, что непонятно сейчас, как можно сделать какие-то осмысленные переводческие диахронические исследования на таких огромных корпусах. Делаются интересные исследования с помощью методов дистрибутивной семантики (метод дистрибутивной семантики показывает контекстуальную близость слов в разные периоды).

С использованием ресурса Google Ngram есть несколько больших проблем, которые на данный момент не решены. Первая проблема связана с тем, что если смотреть, как менялись некие выражения, то на самом деле прослеживается история выражений, история слов, однако ничего неизвестно о том, какие слова употреблялись вместо них. Потому что люди могли просто употреблять совершенно другие конструкции, другие слова, чтобы выражать те же самые смыслы.

Вторая проблема, является более серьезной, состоит в том, что прямым анализом больших данных оказывается весьма сложно отделить лингвистические факторы от экстралингвистических, то есть изменение языка как системы от изменения частотности каких-то слов потому, что изменились реалии.

Самая большая проблема состоит в том, что мы не очень понимаем, как мы можем их анализировать, как мы можем оценить. Тот результат, который мы получили, как оценить, что он валидный, корректный? Над такими методологическими проблемами сейчас идет большая работа.

СПИСОК ЛИТЕРАТУРЫ

1. Гамбье И. Перевод и переводоведение на перекрестке цифровых технологий // Вестник СпбГУ. Серия 9. Филология. Востоковедение. Журналистика. 2016. Вып. 4. С. 56–74.

2. Комиссаров В. Н. Современное переводоведение. Учебное пособие. М.: ЭТС, 2001. 424 с.

3. Петрова О. В., Ланчиков В. К. Сколько гитик умеет перевод. Об одной концепции переводческих стратегий // Мосты. Журнал переводчиков. № 1(53)/2017. М.: Р. Валент, 2017. С.40-50.

4. Pym A. What Technology Doesto Translating // Translation and Interpreting. Vol. 3. No 1 (2011): https://cloud.mail. ru/public/4QBA/ZuUpBXypn/PYM.What technologies does to translating.pdf

5. Прунч Э. Пути развития западного переводоведения. От языковой асимметрии к политической / Пер. с нем. М.: Р. Валент, 2015. 512 с.